PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-228486
(43)Date of publication of application : 25.08.1998

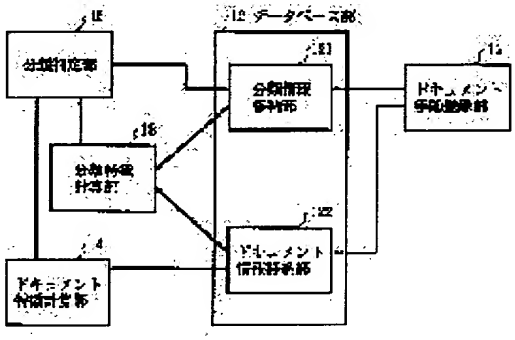
(51)Int.CI. G06F 17/30
G06F 12/00

(21)Application number : 09-047332
(22)Date of filing : 14.02.1997

(71)Applicant : NEC CORP
(72)Inventor : TAKANO HAJIME

(54) DISTRIBUTED DOCUMENT CLASSIFICATION SYSTEM AND RECORDING MEDIUM WHICH RECORDS PROGRAM AND WHICH CAN MECHANICALLY BE READ

(57)Abstract:
PROBLEM TO BE SOLVED: To automatically classify documents and to permit a document generator to designate classification to which the self-document is to belong.



SOLUTION: A service supplier registers information containing a book item characterizing the identifier and the document content of a part of the document in the storage part 122 of a data base part 12, decides a part of classification destinations and registers the identifiers to the classification items of a storage part 121. A calculation part 13 refers to the book item of the classified document in the data base part 12 and calculates the feature amounts of the respective classification items. A calculation part 14 calculates the feature amount of the document which is not classified in the data base part 12 and decides and classifies the classification item to which the document that is not classified is to belong from the calculation result and the feature amounts of the respective classification items. A collection part 66 periodically collects the new or update document from a network environment and a classified information extraction part 67 classifies and registers it to the data base part 12 in accordance with the designation of the classification item in the document.

LEGAL STATUS

[Date of request for examination] 14.02.1997
[Date of sending the examiner's decision of rejection] 09.11.1999
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平10－228486

(43)公開日　平成10年(1998)８月25日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３１０Ｄ
12/00	５４５		12/00	５４５Ｚ
			15/40	３１０Ｃ
				３１０Ｆ
			15/403	３５０Ｃ

審査請求　有　　請求項の数６　ＦＤ　（全 14 頁）

(21)出願番号　特願平9－47332

(22)出願日　平成９年(1997)２月14日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　高野　元
東京都港区芝五丁目７番１号　日本電気株式会社内

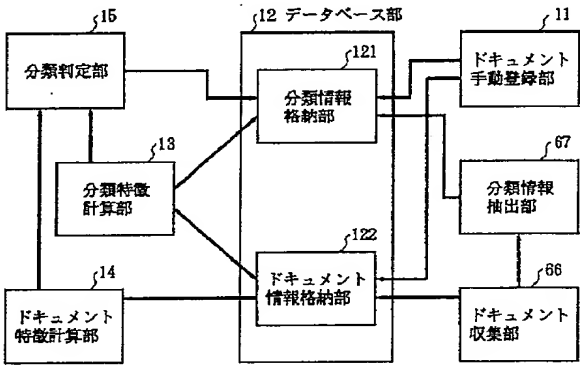
(74)代理人　弁理士　境　廣巳

(54)【発明の名称】　分散ドキュメント分類システム及びプログラムを記録した機械読み取り可能な記録媒体

(57)【要約】

【課題】　ドキュメントの自動分類を可能にし、またドキュメント作成者が自己のドキュメントが属すべき分類を指定できるようにする。

【解決手段】　サービス提供者は一部のドキュメントについて、その識別子及びドキュメント内容を特徴づける書誌項目を含む情報をデータベース部12の格納部122 に登録し、更に一部の分類先を決定して格納部121 の分類項目にそれらの識別子を登録する。計算部13は、データベース部12中の分類済ドキュメントの書誌項目を参照して各分類項目の特徴量を計算する。計算部14は、データベース部12中の未分類ドキュメントの特徴量を計算し、判定部15はその計算結果と各分類項目の特徴量とから未分類ドキュメントが属すべき分類項目を決定して分類する。収集部66は定期的に新規又は更新ドキュメントをネットワーク環境から収集し、分類情報抽出部67はドキュメント中の分類項目指定に従ってデータベース部12に分類登録する。



1

【特許請求の範囲】

【請求項１】　ネットワーク環境に分散して存在するドキュメントを分類する分散ドキュメント分類システムにおいて、

予め指定された分類項目および各分類項目に属するドキュメントの識別子を含む分類情報を格納する分類情報格納部と、各ドキュメントの識別子および各ドキュメントの内容を特徴付ける書誌項目を含むドキュメント情報を格納するドキュメント情報格納部とを有するデータベース部と、

サービス提供者が、ドキュメントのドキュメント情報を前記ドキュメント情報格納部に登録し、且つ、ドキュメントの識別子を前記分類情報格納部の該当する分類項目に登録するためのドキュメント手動登録部と、

前記分類情報格納部に格納された分類情報と前記ドキュメント情報格納部に格納されたドキュメント情報とに基づき、各分類項目の特徴量を計算する分類特徴計算部と、

前記ドキュメント情報格納部にドキュメント情報が登録されているが、そのドキュメント識別子が前記分類情報格納部に登録されていない未分類のドキュメントについて、前記ドキュメント情報格納部に登録されているドキュメント情報に基づき、そのドキュメントの特徴量を計算するドキュメント特徴計算部と、

該ドキュメント特徴計算部で計算された特徴量と前記分類特徴計算部で計算された各分類項目の特徴量とに基づいて、前記未分類のドキュメントが属すべき分類項目を判定し、そのドキュメントの識別子を前記分類情報格納部の前記判定した分類項目に登録する分類判定部とから構成されることを特徴とする分散ドキュメント分類システム。

【請求項２】　新規に作成されたドキュメント及び内容の更新されたドキュメントをネットワーク環境から定期的に収集し、そのドキュメントのドキュメント情報を前記ドキュメント情報格納部に新規登録または更新登録するドキュメント収集部と、

該ドキュメント収集部で収集されたドキュメント中に分類項目指定が存在するか否かを調べ、存在する場合にはそのドキュメントのドキュメント識別子を前記分類情報格納部の前記指定された分類項目に登録する分類情報抽出部とを有することを特徴とする請求項１記載の分散ドキュメント分類システム。

【請求項３】　ネットワーク環境に分散して存在するドキュメントを分類する分散ドキュメント分類システムにおいて、

予め指定された分類項目および各分類項目に属するドキュメントの識別子を含む分類情報を格納する分類情報格納部と、各ドキュメントの識別子および各ドキュメントの内容を特徴付ける書誌項目を含むドキュメント情報を格納するドキュメント情報格納部とを有するデータベース部と、

2

新規に作成されたドキュメント及び内容の更新されたドキュメントをネットワーク環境から定期的に収集し、そのドキュメントのドキュメント情報を前記ドキュメント情報格納部に新規登録または更新登録するドキュメント収集部と、

該ドキュメント収集部で収集されたドキュメント中に分類項目指定が存在するか否かを調べ、存在する場合にはそのドキュメントのドキュメント識別子を前記分類情報格納部の前記指定された分類項目に登録する分類情報抽出部と、

前記分類情報格納部に格納された分類情報と前記ドキュメント情報格納部に格納されたドキュメント情報とに基づき、各分類項目の特徴量を計算する分類特徴計算部と、

前記ドキュメント情報格納部にドキュメント情報が登録されているが、そのドキュメント識別子が前記分類情報格納部に登録されていない未分類のドキュメントについて、前記ドキュメント情報格納部に登録されているドキュメント情報に基づき、そのドキュメントの特徴量を計算するドキュメント特徴計算部と、

該ドキュメント特徴計算部で計算された特徴量と前記分類特徴計算部で計算された各分類項目の特徴量とに基づいて、前記未分類のドキュメントが属すべき分類項目を判定し、そのドキュメントの識別子を前記分類情報格納部の前記判定した分類項目に登録する分類判定部とから構成されることを特徴とする分散ドキュメント分類システム。

【請求項４】　ネットワーク環境に分散して存在するドキュメントを分類するプログラムであって、コンピュータを、

予め指定された分類項目および各分類項目に属するドキュメントの識別子を含む分類情報を格納する分類情報格納部と、各ドキュメントの識別子および各ドキュメントの内容を特徴付ける書誌項目を含むドキュメント情報を格納するドキュメント情報格納部とを有するデータベース部、

サービス提供者が、ドキュメントのドキュメント情報を前記ドキュメント情報格納部に登録し、且つ、ドキュメントの識別子を前記分類情報格納部の該当する分類項目に登録するためのドキュメント手動登録部、

前記分類情報格納部に格納された分類情報と前記ドキュメント情報格納部に格納されたドキュメント情報とに基づき、各分類項目の特徴量を計算する分類特徴計算部、

前記ドキュメント情報格納部にドキュメント情報が登録されているが、そのドキュメント識別子が前記分類情報格納部に登録されていない未分類のドキュメントについて、前記ドキュメント情報格納部に登録されているドキュメント情報に基づき、そのドキュメントの特徴量を計算するドキュメント特徴計算部、

該ドキュメント特徴計算部で計算された特徴量と前記分類特徴計算部で計算された各分類項目の特徴量とに基づいて、前記未分類のドキュメントが属すべき分類項目を判定し、そのドキュメントの識別子を前記分類情報格納部の前記判定した分類項目に登録する分類判定部、
として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【請求項５】　コンピュータを、更に、新規に作成されたドキュメント及び内容の更新されたドキュメントをネットワーク環境から定期的に収集し、そのドキュメントのドキュメント情報を前記ドキュメント情報格納部に新規登録または更新登録するドキュメント収集部、
該ドキュメント収集部で収集されたドキュメント中に分類項目指定が存在するか否かを調べ、存在する場合にはそのドキュメントのドキュメント識別子を前記分類情報格納部の前記指定された分類項目に登録する分類情報抽出部、
として機能させるプログラムを記録した請求項４記載のプログラムを記録した機械読み取り可能な記録媒体。

【請求項６】　ネットワーク環境に分散して存在するドキュメントを分類するプログラムであって、コンピュータを、
予め指定された分類項目および各分類項目に属するドキュメントの識別子を含む分類情報を格納する分類情報格納部と、各ドキュメントの識別子および各ドキュメントの内容を特徴付ける書誌項目を含むドキュメント情報を格納するドキュメント情報格納部とを有するデータベース部、
新規に作成されたドキュメント及び内容の更新されたドキュメントをネットワーク環境から定期的に収集し、そのドキュメントのドキュメント情報を前記ドキュメント情報格納部に新規登録または更新登録するドキュメント収集部、
該ドキュメント収集部で収集されたドキュメント中に分類項目指定が存在するか否かを調べ、存在する場合にはそのドキュメントのドキュメント識別子を前記分類情報格納部の前記指定された分類項目に登録する分類情報抽出部、
前記分類情報格納部に格納された分類情報と前記ドキュメント情報格納部に格納されたドキュメント情報とに基づき、各分類項目の特徴量を計算する分類特徴計算部、
前記ドキュメント情報格納部にドキュメント情報が登録されているが、そのドキュメント識別子が前記分類情報格納部に登録されていない未分類のドキュメントについて、前記ドキュメント情報格納部に登録されているドキュメント情報に基づき、そのドキュメントの特徴量を計算するドキュメント特徴計算部、
該ドキュメント特徴計算部で計算された特徴量と前記分類特徴計算部で計算された各分類項目の特徴量とに基づいて、前記未分類のドキュメントが属すべき分類項目を

判定し、そのドキュメントの識別子を前記分類情報格納部の前記判定した分類項目に登録する分類判定部、
として機能させるプログラムを記録した機械読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明はネットワーク環境に分散して存在するドキュメントを分類するシステムに関し、特に、予め分類項目を用意し、サービス提供者が代表的なドキュメントの内容を判断してこれを分類し、それ以外のドキュメントは既に分類されているドキュメント群との類似度を計算することによって自動的に分類する分散ドキュメント分類システムに関する。

【０００２】

【従来の技術】Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ（以下、WWWと称す）のように多数のドキュメントが分散して存在するネットワーク環境においては、ドキュメント数に比例して所望のドキュメントを探し出すことが困難になる。このため、WWWのディレクトリサービスのように、ネットワーク上に分散するドキュメントを予め分類してその所在や書誌項目をデータベースに蓄積し、クライアントに対して検索サービスを提供するディレクトリサービスが普及している。本発明はこのようなディレクトリサービスなどを実現する際に必要な分散ドキュメント分類システムに関する。

【０００３】図１４にこの種の従来の分散ドキュメント分類システムの構成を示す。同図に示すように従来の分散ドキュメント分類システムは、分類情報格納部９２１およびドキュメント情報格納部９２２を含むデータベース部９２と、ドキュメント手動登録部９１とからなる。

【０００４】データベース部９２において、ドキュメント情報格納部９２２は、ネットワーク上に分散して存在するドキュメントのドキュメント識別子とその書誌項目のリストとを保持する部分であり、分類情報格納部９２１は、分類項目のリストと各分類項目に分類されているドキュメントのドキュメント識別子とを保持する部分である。サービス管理者は、新たなドキュメントを登録する場合、そのドキュメントの内容を確認して書誌項目として使用すべき項目を判断し、この判断した書誌項目と何らかの方法で決定したドキュメント識別子とを、ドキュメント手動登録部９１を通じてドキュメント情報格納部９２２に追加し、また、上記確認したドキュメントの内容から当該ドキュメントが分類される項目を判断し、分類情報格納部９２１中の該当する分類項目に当該ドキュメントのドキュメント識別子を追加登録する。

【０００５】上記の分散ドキュメント分類システムは、ドキュメントの分類作業を全て人手で行うシステムであるが、それを自動的に行うシステムも提案されている。例えば、特開平７－４９８７５号公報では、各分類ごとに予め用意した検索条件としての単語リストと各ドキュ

メント間の適合度を計算することによって、自動的にドキュメントを分類している。また、ネットワーク上のドキュメントの更新状況を監視して、更新があったドキュメントを収集し、分類処理するようにしている。

【０００６】

【発明が解決しようとする課題】しかしながら、図１４に示されるような従来の分散ドキュメント分類システムでは、ドキュメント識別子や書誌項目の登録ならびに各ドキュメントの分類作業は、ドキュメント手動登録部を通じて全てサービス提供者が行う必要があるため、非常にコストがかかるという問題点があった。

【０００７】他方、特開平７－４９８７５号公報記載のシステムでは、文書の分類を自動的に行うことができる。しかし、全ての文書を自動分類することを前提としているため、各分類に付与する検索条件を前もって設定しておく必要がある。検索条件は単語のリストなどであるが、一つの文書も分類していない状態で、各分類ごとの検索条件を適切に設定することは相当なスキルが要求される。また、或る分類項目に分類されるべき幾つかの文書を実際に調べて検索条件を求める作業を行うとしても、その作業に使用した文書は既に分類先が決まっているにもかかわらず、これらも自動分類の対象としなければならず、無駄が多い。

【０００８】したがって本発明の第１の目的は、手動分類と自動分類とを組み合わせ、ネットワーク環境に分散して存在するドキュメントの一部を手動で分類する作業をサービス提供者が行えば、その他のドキュメントは既に分類されているドキュメント群との類似度を計算して自動的に分類する分散ドキュメント分類システムを提供することにある。

【０００９】また、サービス提供者によるドキュメントの手動分類やこの手動分類されたドキュメント群との類似度による他のドキュメントの自動分類では、ドキュメントの分類結果は、分類を行うサービス提供者の判断に依存する部分が多く、必ずしもドキュメント作成者の意図と一致するとは限らない。特開平７－４９８７５号公報に記載された技術では、そもそも全ての文書を検索条件に基づいて自動的に分類してしまうため、文書作成者の意図とは無関係に分類されてしまう。ドキュメントを作成した者はそのドキュメントに関して最も熟知している者であるため、このようなドキュメント作成者の協力があれば、より一層適切な分類が可能になるであろう。

【００１０】したがって本発明の第２の目的は、ドキュメント作成者が自己のドキュメントが属すべき分類を明示的に指定することができる分散ドキュメント分類システムを提供することにある。

【００１１】

【課題を解決するための手段】本発明は上記第１の目的を達成するために、ネットワーク環境に分散して存在するドキュメントを分類する分散ドキュメント分類システムにおいて、予め指定された分類項目および各分類項目に属するドキュメントの識別子を含む分類情報を格納する分類情報格納部と、各ドキュメントの識別子および各ドキュメントの内容を特徴付ける書誌項目を含むドキュメント情報を格納するドキュメント情報格納部とを有するデータベース部と、サービス提供者が、ドキュメントのドキュメント情報を前記ドキュメント情報格納部に登録し、且つ、ドキュメントの識別子を前記分類情報格納部の該当する分類項目に登録するためのドキュメント手動登録部と、前記分類情報格納部に格納された分類情報と前記ドキュメント情報格納部に格納されたドキュメント情報とに基づき、各分類項目の特徴量を計算する分類特徴計算部と、前記ドキュメント情報格納部にドキュメント情報が登録されているが、そのドキュメント識別子が前記分類情報格納部に登録されていない未分類のドキュメントについて、前記ドキュメント情報格納部に登録されているドキュメント情報に基づき、そのドキュメントの特徴量を計算するドキュメント特徴計算部と、該ドキュメント特徴計算部で計算された特徴量と前記分類特徴計算部で計算された各分類項目の特徴量とに基づいて、前記未分類のドキュメントが属すべき分類項目を判定し、そのドキュメントの識別子を前記分類情報格納部の前記判定した分類項目に登録する分類判定部とから構成される。

【００１２】このように構成された本発明の分散ドキュメント分類システムにあっては、サービス提供者は、ネットワーク環境に分散して存在するドキュメントの内容から、書誌項目として使用すべき項目を判断し、ドキュメント手動登録部を通じて、この判断した書誌項目とそのドキュメントを一意に識別するドキュメント識別子とを含むドキュメント情報をデータベース部のドキュメント情報格納部に登録し、また、幾つかの代表的なドキュメントについてその内容から当該ドキュメントが分類される項目を判断し、分類情報格納部の該当する分類項目に当該ドキュメントのドキュメント識別子を登録しておく。こうしておくと、後は、ドキュメント情報格納部にドキュメント情報が格納されている未分類のドキュメントが以下のように自動的に分類される。まず、分類特徴計算部が、分類情報格納部に格納された分類情報とドキュメント情報格納部に格納されたドキュメント情報とに基づき、各分類項目の特徴量を計算し、ドキュメント特徴計算部が、ドキュメント情報格納部にドキュメント情報が登録されているが、そのドキュメント識別子が分類情報格納部に登録されていない未分類のドキュメントについて、ドキュメント情報格納部に登録されているドキュメント情報に基づき、そのドキュメントの特徴量を計算する。そして、分類判定部が、ドキュメント特徴計算部で計算された特徴量と分類特徴計算部で計算された各分類項目の特徴量とに基づいて、前記未分類のドキュメントが属すべき分類項目を判定し、そのドキュメントの

識別子を分類情報格納部の前記判定した分類項目に登録する。

【００１３】また本発明は上記第１および第２の目的をも達成するために、更に、新規に作成されたドキュメント及び内容の更新されたドキュメントをネットワーク環境から定期的に収集し、そのドキュメントのドキュメント情報を前記ドキュメント情報格納部に新規登録または更新登録するドキュメント収集部と、該ドキュメント収集部で収集されたドキュメント中に分類項目指定が存在するか否かを調べ、存在する場合にはそのドキュメントのドキュメント識別子を前記分類情報格納部の前記指定された分類項目に登録する分類情報抽出部とを有する。

【００１４】このような構成の分散ドキュメント分類システムにあっては、ドキュメントの新規作成時または更新時に、ドキュメント作成者がドキュメント中に分類項目指定を記述しておくと、ドキュメント収集部がこのドキュメントを収集した際に、そのドキュメントのドキュメント情報をドキュメント情報格納部に新規登録または更新登録し、分類情報抽出部が、収集されたドキュメント中に記述された分類項目指定に従ってそのドキュメントのドキュメント識別子を分類情報格納部の前記指定された分類項目に登録することで、ドキュメント作成者の意図する分類項目に分類する。

【００１５】以上の構成では、サービス提供者がドキュメント手動登録部を使用して一部のドキュメントを手動登録することを前提としたが、ドキュメント作成者の比較的多くの者が自己のドキュメントに分類項目指定を含めるような状況下では、各分類項目の特徴量を求める基礎となる分類済みドキュメントとして、ドキュメント作成者が分類指定したドキュメントだけで足りるようになり、その場合には、サービス提供者による手動分類は一切不要で、ドキュメント手動登録部を省略することができる。かかる構成の分散ドキュメント分類システムは、上述した、データベース部と、ドキュメント収集部と、分類情報抽出部と、分類特徴計算部と、ドキュメント特徴計算部と、分類判定部とから構成される。動作にあっては、１回目の起動時に、ドキュメント収集部がネットワーク環境に分散するドキュメントを収集して、そのドキュメント情報をデータベース部のドキュメント情報格納部に格納し、次いで、収集されたドキュメントのうち分類項目指定のあるドキュメントを分類情報抽出部によって分類する。この後、分類特徴計算部が、既分類のドキュメントの書誌項目に従って各分類項目の特徴量を計算し、ドキュメント特徴計算部が未分類のドキュメントの特徴量を計算し、分類判定部がそれらの結果から未分類のドキュメントの属すべき分類項目を決定して分類する。次回の定期起動時には、新規または更新ドキュメントだけが収集され、そのうち分類項目指定のあるものは分類情報抽出部で分類登録され、分類項目指定のないものは分類判定部によって自動的に分類される。

【００１６】

【発明の実施の形態】次に本発明の実施の形態の例について図面を参照して詳細に説明する。

【００１７】図１は本発明の第１の実施例にかかる分散ドキュメント分類システムの構成を示すブロック図である。

【００１８】図示のように、本実施例の分散ドキュメント分類システムは、ドキュメント手動登録部１１と、データベース部１２と、分類特徴計算部１３と、ドキュメント特徴計算部１４と、分類判定部１５とから構成されている。なお、図には本実施例の特徴的な構成のみが記載されており、他の構成は省略してある。

【００１９】データベース部１２は、ネットワーク環境に分散して存在するドキュメントを分類して登録しておく部分であり、分類情報格納部１２１とドキュメント情報格納部１２２とを備えている。

【００２０】分類情報格納部１２１は、ドキュメントの分類項目および各分類項目に属するドキュメントの識別子を含む分類情報を格納する部分である。図２は、分類情報格納部１２１に格納される情報のデータ構造の一例を示すテーブルである。図示のように、テーブル２０は、カラムとして、分類ＩＤと、分類項目と、その分類に属しているドキュメントの識別子のリストとを格納する行（レコード）を、複数行有している。

【００２１】ドキュメント情報格納部１２２は、各ドキュメントの識別子および各ドキュメントの内容を特徴づける書誌項目などを含むドキュメント情報を格納する部分である。図３は、ドキュメント情報格納部１２２に格納される情報のデータ構造の一例を示すテーブルである。図示のように、テーブル３０は、カラムとして、ドキュメント識別子と、当該ドキュメントのネットワーク上のアドレスと、当該ドキュメントの書誌項目と、分類方法とを持つ行（レコード）を、複数行有している。

【００２２】以上のようなデータベース部１２は、例えばプログラム制御されたＣＰＵおよび磁気ディスク装置などの記憶装置で実現される。

【００２３】ドキュメント手動登録部１１は、ドキュメントを手動で分類して登録する際に使用する部分である。サービス提供者は、ドキュメント手動登録部１１を通じて、ネットワーク環境に分散して存在するドキュメントについて、そのドキュメント識別子と書誌項目などを含むドキュメント情報をデータベース部１２のドキュメント情報格納部１２２に登録し、また、各分類項目に属すべき一部のドキュメントについて、そのドキュメント識別子を分類情報格納部１２１中の該当する分類項目に登録する。更に、分類情報格納部１２１に分類項目を定義する場合にもこのドキュメント手動登録部１１が使用される。このようなドキュメント手動登録部１１は、例えばプログラム制御されたＣＰＵを用いた制御手段およびディスプレイ装置などで実現される。

【００２４】分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５は、手動で分類しなかった残りのドキュメントを自動分類するためのものである。このうち、分類特徴計算部１３は、データベース部１２中で既に各分類項目に分類されているドキュメントの書誌項目を参照して、各分類項目の特徴量を計算する部分である。また、ドキュメント特徴計算部１４は、データベース部１２中の未分類のドキュメントの特徴量をその書誌項目に基づいて計算する部分である。更に、分類判定部１５は、ドキュメント特徴計算部１４で計算された特徴量と分類特徴計算部１３で計算された各分類項目の特徴量とに基づいて、未分類のドキュメントが属すべき分類項目を判定し、分類登録する部分である。

【００２５】これらの分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５は、例えばプログラム制御されたＣＰＵおよびメモリなどの一時記憶装置で実現される。

【００２６】次に本実施例の各部の詳細な機能を、全体の動作と共に説明する。

【００２７】サービス提供者は、ネットワーク環境に分散して存在するドキュメントに適用する分類項目を予め用意しておき、これをドキュメント手動登録部１１を通じてデータベース部１２の分類情報格納部１２１に定義する。この定義にしたがって、分類項目の識別子が図２のテーブル２０のカラム「分類ＩＤ」に、分類名がカラム「分類」に格納される。なお、分類項目の識別子は任意のもので良いが、ここでは０から始まる連続番号を付与している。

【００２８】また、サービス提供者は分類対象となるドキュメントに関するドキュメント情報を、ドキュメント手動登録部１１を通じてデータベース部１２のドキュメント情報格納部１２２に設定する。設定するドキュメント情報としては、ドキュメント識別子，そのドキュメントのネットワーク上の所在を示すアドレス，そのドキュメントの内容を特徴づける書誌項目などである。この設定操作にしたがって、各ドキュメント毎に図３のテーブル３０の１行が割り当てられ、その行のカラム「ドキュメント識別子」，カラム「アドレス」，カラム「書誌項目」にそれぞれドキュメント識別子，アドレス，書誌項目が格納される。なお、カラム「分類方法」はこの時点ではＮＵＬＬである。

【００２９】さらにサービス提供者は、すでにドキュメント情報格納部１２２にドキュメント情報が登録されているドキュメントのうち、各分類項目に属すべき代表的なドキュメントの内容を判別し、ドキュメント手動登録部１１を通じて適切な分類項目を指定する。この指定にしたがって、ドキュメント手動登録部１１は、指定されたドキュメントのドキュメント識別子D_iを、分類情報格納部１２１における指定された分類項目のドキュメント識別子のリストに追加する。同時に、ドキュメント手動登録部１１は、ドキュメント情報格納部１２２におけるその指定されたドキュメント識別子D_iを持つ行のカラム「分類方法」に、手動で分類したことを示す値“ｍａｎｕａｌ”を設定する。なお、ドキュメント情報格納部１２２に登録したドキュメント全てを手動で分類する必要はないが、少なくとも各分類項目には１つ以上のドキュメントを分類しておく必要があり、また、後述する分類項目の特徴量の精度を上げるためには、ある程度の数のドキュメントを各分類項目に分類しておく必要がある。なお、手動で分類されなかった残りのドキュメントのドキュメント情報格納部１２２におけるカラム「分類方法」の値はＮＵＬＬのままである。

【００３０】以上に述べたような登録作業に引き続き、分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５による自動分類処理が行われる。

【００３１】まず、分類特徴計算部１３は、分類情報格納部１２１に格納された各分類項目ごとにその特徴量を計算する。図４に特徴量計算の手順例を示す。

【００３２】分類特徴計算部１３は、まず、読み出しカウンタｉを０にリセットし（ｓｔｅｐ４１）、分類情報格納部１２１から分類識別子C_iの行のデータを読み出す（ｓｔｅｐ４２）。次に、読み出したデータに含まれるドキュメント識別子のリストにしたがって、ドキュメント情報格納部１２２からその分類項目C_iに分類されている全てのドキュメントの書誌項目を読み出し、一時記憶に格納する（ｓｔｅｐ４４）。次に、一時記憶に格納した書誌項目に基づいて分類項目C_iの特徴量K_iを計算し（ｓｔｅｐ４５）、これを内部の特徴量記憶部に格納する（ｓｔｅｐ４６）。そして、読み出しカウンタｉを＋１し（ｓｔｅｐ４７）、ｓｔｅｐ４２に戻って上述した処理を繰り返す。ｓｔｅｐ４３で分類情報格納部１２１から全ての分類項目のデータを読み出し終えたことを検出した時点で、処理を終了する。

【００３３】分類特徴計算部１３において各分類項目の特徴量の計算が終了すると、次にドキュメント特徴計算部１４が未分類のドキュメントの特徴量を計算し、この値をもとに分類判定部１５がそのドキュメントが属すべき分類項目を判定し、登録を行う。図５に、ドキュメント特徴計算部１４および分類判定部１５の処理の手順例を示す。

【００３４】まず、ドキュメント特徴計算部１４は、ドキュメント情報格納部１２２からまだ分類されていない、つまりカラム「分類方法」がＮＵＬＬのドキュメント（ドキュメント識別子をD_iとする）の行のデータを一つ読み出し（ｓｔｅｐ５１）、そのデータに記述されている書誌項目から、当該ドキュメントの特徴量DK_iを計算する（ｓｔｅｐ５３）。

【００３５】続いて、分類判定部１５は、この計算されたドキュメントの特徴量DK_iと、分類特徴計算部１３内部の特徴量記憶部に格納されている各分類項目の特徴

量K_i（ｉ＝０，１，…）とから、当該ドキュメントの属すべき分類項目Ｃ。を選択する（ｓｔｅｐ５４）。この選択では、ドキュメントの特徴量DK_iに最も値が近い特徴量K_iを持つ１つの分類項目を選択しても良く、その差が或る閾値の範囲内に収まる１つ以上の分類項目を選択しても良い。

【００３６】次に分類判定部１５は、分類情報格納部１２１に格納されている上記選択した分類項目Ｃ。のドキュメント識別子のリストに、当該ドキュメントの識別子D_iを追加し（ｓｔｅｐ５５）、さらにドキュメント情報格納部１２２に格納されている当該ドキュメント識別子D_iを持つ行のカラム「分類方法」に、自動で分類したことを示す“ａｕｔｏ”を設定する（ｓｔｅｐ５６）。

【００３７】そして、ｓｔｅｐ５１に戻って上述した処理を繰り返し、ｓｔｅｐ５２でドキュメント情報格納部１２２から未分類の全てのドキュメントを読み出し終えたことを検出した時点で、処理を終了する。

【００３８】以上のような自動分類処理が行われることにより、ドキュメント情報格納部１２２に格納された全てのドキュメントの分類が終了する。

【００３９】次に、書誌項目の具体例を示し、それに基づいて分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５の動作をより具体的に説明する。

【００４０】ネットワーク環境に分散して存在するドキュメントは、画像や音声等のマルチメディアデータであって良いが、ここでは、テキストを含む文書とする。そして、ドキュメントの内容を特徴づける書誌項目として、文書中に一定回数以上（例えば２回以上）出現する単語（キーワード）とその出現回数の組の集合とする。例えば或るドキュメントＸ中に、図６（ａ）に示すように、「テニス」，「スキー」，「サッカー」という単語がそれぞれ５回，４回，３回出現しており、それ以外に２回以上出現する単語がないものとすると、「テニス」＝５回，「スキー」＝４回，「サッカー」＝３回がドキュメントＸの書誌項目となる。同様に別のドキュメントＹ中に、図６（ｂ）に示すように、「テニス」，「野球」，「スキー」という単語がそれぞれ３回，２回，６回出現しており、それ以外に２回以上出現する単語がないものとすると、「テニス」＝３回，「野球」＝２回，「スキー」＝６回がドキュメントＹの書誌項目となる。

【００４１】ドキュメント特徴計算部１４は、ドキュメントの書誌項目に記述された各単語の出現回数を全単語の総出現回数で割った出現頻度の集合を、当該ドキュメントの特徴量とする。従って、ドキュメントＸの場合には図６（ａ）に示すように、「テニス」＝５／１２，「スキー」＝４／１２，「サッカー」＝３／１２が特徴量となり、ドキュメントＹの場合には図６（ｂ）に示すように、「テニス」＝３／１１，「野球」＝２／１１，「スキー」＝６／１１が特徴量となる。

【００４２】また分類特徴計算部１３は、各分類項目ごとに、その分類項目に属するドキュメントの書誌項目に現れる各単語の出現回数の総和を各単語ごとにカウントし、これを全単語の総出現回数で割った出現頻度の集合を、その分類項目の特徴量とする。例えば、ドキュメントＸとドキュメントＹとが同じ分類項目に属するものとし、その分類項目にはそれ以外のドキュメントが分類されていないとすると、図６（ａ），（ｂ）の内容から、図７に示すように、単語「テニス」，「スキー」，「サッカー」，「野球」の総出現回数がそれぞれ８，１０，３，２として求まり、全単語の総出現回数は２３なので、図７に示すように、「テニス」＝８／２３，「スキー」＝１０／２３，「サッカー」＝３／２３，「野球」＝２／２３が当該分類項目の特徴量となる。

【００４３】分類判定部１５は、ドキュメント特徴計算部１４で計算された未分類のドキュメントの特徴量に現れる単語の出現頻度と、分類特徴計算部１３で計算された各分類項目の特徴量に現れるこれと同一単語の出現頻度との積の総和を類似度とする。例えば、未分類のドキュメントＺの書誌項目が図８に示すように、「テニス」＝４，「スキー」＝２，「ゴルフ」＝３であった場合、その特徴量は同図に示すように「テニス」＝４／９，「スキー」＝２／９，「ゴルフ」＝３／９となる。従って、このドキュメントＺと図７に示した分類項目との類似度は、（４／９）×（８／２３）＋（２／９）×（１０／２３）として求められる。

【００４４】図９は本発明の第２の実施例にかかる分散ドキュメント分類システムの構成を示すブロック図であり、図１と同一符号は同一部分を示し、６６はドキュメント収集部、６７は分類情報抽出部である。

【００４５】本実施例の分散ドキュメント分類システムは、ドキュメント収集部６６および分類情報抽出部６７を更に有する点で図１に示した実施例の分散ドキュメント分類システムと相違する。

【００４６】ドキュメント手動登録部１１，データベース部１２，分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５は、基本的に図１の実施例のものと同じである。

【００４７】ドキュメント収集部６６は、データベース部１２のドキュメント情報格納部１２２に登録されていない新規なドキュメント、および登録されているがその内容が更新されたドキュメントを、ネットワーク環境から定期的に収集し、そのドキュメント情報をドキュメント情報格納部１２２に新規登録または更新登録する部分である。

【００４８】分類情報抽出部６７は、ドキュメント収集部６６で収集されたドキュメント中にドキュメント作成者が記述した分類項目指定が存在するか否かを調べ、存在する場合にはそのドキュメントのドキュメント識別子

を分類情報格納部１２１の前記指定された分類項目に登録する部分である。

【００４９】これらのドキュメント収集部６６および分類情報抽出部６７は、例えばプログラム制御されたＣＰＵおよびメモリや磁気ディスクなどの記憶装置で実現される。

【００５０】次に本実施例の動作を、第１の実施例と相違する部分を中心に説明する。

【００５１】サービス提供者は、ネットワーク環境に分散して存在するドキュメントに適用する分類項目を予め用意しておき、これをドキュメント手動登録部１１を通じてデータベース部１２の分類情報格納部１２１に定義する。この定義にしたがって、分類項目の識別子が図２のテーブル２０のカラム「分類ＩＤ」に、分類名がカラム「分類」に格納される。

【００５２】また、サービス提供者は分類対象となるドキュメントに関するドキュメント情報を、ドキュメント手動登録部１１を通じてデータベース部１２のドキュメント情報格納部１２２に設定する。本実施例の場合、ドキュメント収集部６６によって後述する定期的なドキュメント収集が行われるため、サービス提供者は分類対象となる全ドキュメントに関するドキュメント情報を必ずしも登録する必要はない。設定するドキュメント情報としては、ドキュメント識別子，そのドキュメントのネットワーク上の所在を示すアドレス，そのドキュメントの内容を特徴づける書誌項目などである。書誌項目の具体的な例は第１の実施例と同じである。この設定操作にしたがって、各ドキュメント毎に図３のテーブル３０の１行が割り当てられ、その行のカラム「ドキュメント識別子」，カラム「アドレス」，カラム「書誌項目」にそれぞれドキュメント識別子，アドレス，書誌項目が格納される。なお、カラム「分類方法」はこの時点ではＮＵＬＬである。

【００５３】さらにサービス提供者は、すでにドキュメント情報格納部１２２にドキュメント情報が登録されているドキュメントのうち、代表的なドキュメントの内容を判別し、ドキュメント手動登録部１１を通じて適切な分類項目を指定する。この指定にしたがって、ドキュメント手動登録部１１は、指定されたドキュメントのドキュメント識別子D_iを、分類情報格納部１２１における指定された分類名のドキュメント識別子のリストに追加する。同時に、ドキュメント手動登録部１１は、ドキュメント情報格納部１２２におけるその指定されたドキュメント識別子D_iを持つ行のカラム「分類方法」に、手動で分類したことを示す値“ｍａｎｕａｌ”を設定する。なお、ドキュメント情報格納部１２２に登録したドキュメント全てを手動で分類する必要はないが、本動作例では少なくとも各分類項目には１つ以上のドキュメントを分類しておく必要があり、また、後述する分類項目の特徴量の精度を上げるためには、ある程度の数のドキュメントを各分類項目に分類しておく必要がある。なお、手動で分類されなかった残りのドキュメントのドキュメント情報格納部１２２におけるカラム「分類方法」の値はＮＵＬＬのままである。

【００５４】また、サービス提供者は、ドキュメント手動登録部１１を通じてドキュメント情報格納部１２２に登録した各ドキュメントごとに、そのネットワーク上の所在を示すアドレスとそのドキュメントの更新日時との組をファイルに記録しておく。このファイルはドキュメント収集部６６によって参照される。

【００５５】以上に述べたような登録作業に引き続き、分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５による自動分類処理が、第１の実施例と同様に行われる。これにより、サービス提供者がドキュメント情報格納部１２２に格納した全てのドキュメントの分類が終了する。

【００５６】続いて、ドキュメント収集部６６および分類情報抽出部６７が例えば１日単位や１週単位といった周期で定期的に起動される。

【００５７】図１０はドキュメント収集部６６および分類情報抽出部６７の処理例を示すフローチャートである。ドキュメント収集部６６は起動されると、ネットワーク環境から新規または更新されたドキュメントのアドレスを取得する（ｓｔｅｐ８１）。これは、例えば以下のようにして行う。まず、ネットワークを介して各種のサーバにアクセスして、ネットワーク環境に分散して現に存在するドキュメントのアドレスとその更新日時とを収集し、ファイルに記録する。次に、前回の起動時に同様にしてファイルに記録していたアドレスと更新日時との集合と、今回ファイルに記録したアドレスと更新日時との集合を照合する。なお、１回目の起動時には、前回の記録として前述したサービス提供者が作成したファイルを使用する。そして、前回のファイルに記録されておらず、今回のファイルに記録されているアドレスを、新規ドキュメントのアドレスとして抽出する。また、今回のファイルにも、前回のファイルにも同じアドレスが記録されているアドレスについては、その両者の更新日時を比較し、相違するアドレスを、更新ドキュメントのアドレスとして抽出する。そして、それらのアドレスを一時記憶に格納する。この一時記憶に格納されたアドレスが、ネットワーク環境に新規に存在したドキュメントまたは更新されたドキュメントのアドレスとなる。

【００５８】次にドキュメント収集部６６は、一時記憶から１つのアドレスを読み出し（ｓｔｅｐ８２）、そのアドレスに従ってドキュメントをダウンロードする（ｓｔｅｐ８４）。そして、そのドキュメントの内容から書誌項目を決定し、この決定した書誌項目，アドレスおよびドキュメント識別子を含むドキュメント情報をドキュメント情報格納部１２２に登録する（ｓｔｅｐ８５）。このとき、同じアドレスを持つドキュメント情報が既に

ドキュメント情報格納部１２２に登録されている場合、そのドキュメント情報を削除すると共に、この削除したドキュメント情報のドキュメント識別子を分類情報格納部１２１から削除する。

【００５９】続いて、ダウンロードされたドキュメントの内容およびそれに付与されたドキュメント識別子がドキュメント収集部６６から分類情報抽出部６７に伝達され、分類情報抽出部６７は、そのドキュメントの内容に分類項目指定が存在するか否かを調べる（ｓｔｅｐ８６）。図１１に、ドキュメント作成者がドキュメント中に記述する分類項目指定の例を示す。図１１に示すように分類項目指定の記述は、ヘッダ７１とコマンド名７２と引数７３とから構成される。ヘッダ７１は分類機能の動作記述であることを示し、コマンド名７２は分類項目を指定するものであることを示し、引数７３は分類項目として分類１を指定することを示している。

【００６０】分類情報抽出部６７は、ドキュメント中に分類項目指定が存在しない場合は、ｓｔｅｐ８２に戻って上述した処理を繰り返す。また、ドキュメント中に分類項目指定が存在した場合は、指定された分類項目を抽出し（ｓｔｅｐ８７）、分類情報格納部１２１における指定された分類項目のドキュメント識別子のリストに、当該ドキュメントのドキュメント識別子を追加する（ｓｔｅｐ８８）。また、当該ドキュメント識別子を含む、ドキュメント情報格納部１２２中の行のカラム「分類方法」に、ドキュメント作成者の指定による分類であることを示す値“ｓｐｅｃｉｆｙ”を設定する（ｓｔｅｐ８９）。そして、ｓｔｅｐ８２に戻って上述した処理を繰り返す。

【００６１】以上のような処理が繰り返され、新規または更新ドキュメントの全アドレスについての処理を終えたことをｓｔｅｐ８３で検出すると、今回の処理を終了する。これにより、ドキュメント収集部６６で収集された新規または更新ドキュメントのうち、ドキュメント作成者による分類項目指定が存在するドキュメントは、その指定に従って分類されたことになる。

【００６２】その後、サービス提供者は、ドキュメント特徴計算部１４および分類判定部１５に起動をかけ、ドキュメント収集部６６によって収集されたが分類情報抽出部６７によっては分類されなかったドキュメント、つまり分類項目指定のなかったドキュメントの自動分類を行う。なお、分類情報抽出部６７が図１０のｓｔｅｐ８６で分類項目指定のなかったドキュメントが存在したことを検出した場合に、処理の終了後に分類情報抽出部６７からドキュメント特徴計算部１４および分類判定部１５を起動するようにしても良い。ドキュメント特徴計算部１４および分類判定部１５は起動をかけられると、前述と同様に図５に示す処理を行うことにより、ドキュメント情報格納部１２２に格納された未分類のドキュメントを自動的に分類する。

【００６３】また、この例では、分類特徴計算部１３は起動しなかったので、各分類項目の特徴量は、サービス提供者が分類登録したドキュメントに基づいて先に決定した値が使用される。別の実施例として、分類特徴計算部１３も起動し、各分類項目の特徴量を再計算させても良い。こうすると、ドキュメント作成者の記述した分類項目指定によって分類されたドキュメントの書誌項目をも考慮して、各分類項目の特徴量が求められることになる。

【００６４】さらに以上の動作例では、サービス提供者がドキュメントをデータベース部１２に登録して一部のドキュメントを分類し、次いで、分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５によって、サービス提供者が分類しなかったドキュメントを分類し、その後に、ドキュメント収集部６６および分類情報抽出部６７の１回目の起動を行って新規および更新ドキュメントのデータベース部１２への登録と分類項目指定のあるドキュメントの分類とを行い、そして、再びドキュメント特徴計算部１４および分類判定部１５による未分類のドキュメントの分類処理を行わせた。しかし、他の実施例として、サービス提供者がドキュメントをデータベース部１２に登録して一部のドキュメントを分類し、次いで、ドキュメント収集部６６および分類情報抽出部６７の１回目の起動を行って新規および更新ドキュメントのデータベース部１２への登録と分類項目指定のあるドキュメントの分類とを行い、次いで、分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５によって、未分類のドキュメント（サービス提供者が分類しなかったドキュメント及び分類項目指定のなかったドキュメント）を分類するようにしても良い。

【００６５】図１２は本発明の第３の実施例にかかる分散ドキュメント分類システムの構成を示すブロック図であり、図９と同一符号は同一部分を示す。

【００６６】本実施例の分散ドキュメント分類システムは、ドキュメント手動登録部１１を有していない点で図９に示した第２の実施例の分散ドキュメント分類システムと相違する。

【００６７】データベース部１２，分類特徴計算部１３，ドキュメント特徴計算部１４，分類判定部１５，ドキュメント収集部６６および分類情報抽出部６７は、基本的に図９の第２の実施例のものと同じである。

【００６８】次に本実施例の動作を、第２の実施例と相違する部分を中心に説明する。

【００６９】サービス提供者は、ドキュメント収集部６６および分類情報抽出部６７を例えば１日単位や１週単位といった周期で定期的に起動する。勿論、これらを定期的に自動で起動する仕組みを組み込んでも良い。

【００７０】ドキュメント収集部６６および分類情報抽出部６７は起動されると、図１０に示した処理を開始す

る。まず、ドキュメント収集部６６は、ネットワーク環境から新規または更新されたドキュメントのアドレスを取得し、一時記憶に記録する（ｓｔｅｐ８１）。この取得は第２の実施例と同様に行われる。但し、１回目の気時には前回ファイルが存在しないため、今回のファイルに記録されたアドレス全てが一時記憶に移される。次にドキュメント収集部６６は、一時記憶から１つのアドレスを読み出し（ｓｔｅｐ８２）、そのアドレスに従ってドキュメントをダウンロードし（ｓｔｅｐ８４）、そのドキュメントの内容から書誌項目を決定し、この決定した書誌項目，アドレスおよびドキュメント識別子を含むドキュメント情報をドキュメント情報格納部１２２に登録する（ｓｔｅｐ８５）。このとき、同じアドレスを持つドキュメント情報が既にドキュメント情報格納部１２２に登録されている場合、そのドキュメント情報を削除すると共に、この削除したドキュメント情報のドキュメント識別子を分類情報格納部１２１から削除する。

【００７１】続いて、ダウンロードされたドキュメントの内容およびそれに付与されたドキュメント識別子がドキュメント収集部６６から分類情報抽出部６７に伝達され、分類情報抽出部６７は、そのドキュメントの内容に図１１に例示したような分類項目指定が存在するか否かを調べる（ｓｔｅｐ８６）。ドキュメント中に分類項目指定が存在しない場合は、ｓｔｅｐ８２に戻って上述した処理を繰り返す。また、ドキュメント中に分類項目指定が存在した場合は、指定された分類項目を抽出し（ｓｔｅｐ８７）、分類情報格納部１２１における指定された分類項目のドキュメント識別子のリストに、当該ドキュメントのドキュメント識別子を追加する（ｓｔｅｐ８８）。また、当該ドキュメント識別子を含む、ドキュメント情報格納部１２２中の行のカラム「分類方法」に、ドキュメント作成者の指定による分類であることを示す値“ｓｐｅｃｉｆｙ”を設定する（ｓｔｅｐ８９）。そして、ｓｔｅｐ８２に戻って上述した処理を繰り返す。

【００７２】以上のような処理が繰り返され、新規または更新ドキュメントの全アドレスについての処理を終えたことをｓｔｅｐ８３で検出すると、今回の処理を終了する。これにより、ドキュメント収集部６６で収集された新規または更新ドキュメントのうち、ドキュメント作成者による分類項目指定が存在するドキュメントは、その指定に従って分類される。

【００７３】その後、サービス提供者は、分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５に起動をかけ、ドキュメント収集部６６によって収集されたが分類情報抽出部６７によっては分類されなかったドキュメント、つまり分類項目指定のなかったドキュメントの自動分類を行う。なお、分類情報抽出部６７が図１０のｓｔｅｐ８６で分類項目指定のなかったドキュメントが存在したことを検出した場合に、処理の終了後に分類情報抽出部６７から分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５を起動するようにしても良い。分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５は起動をかけられると、第１および第２の実施例と同様に図４および図５に示す処理を行うことにより、ドキュメント情報格納部１２２に格納された未分類のドキュメントを自動的に分類する。

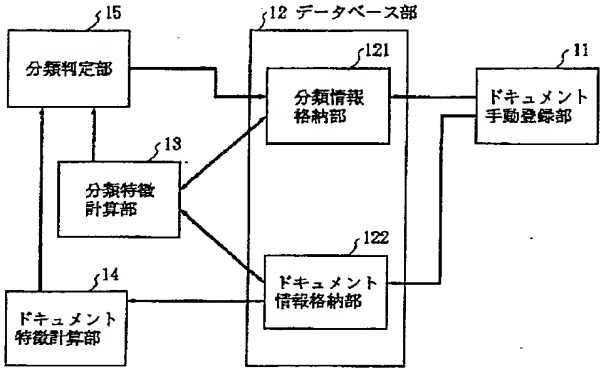
【００７４】ドキュメント収集部６６および分類情報抽出部６７は、１日後あるいは１週間後に再び起動され、ネットワーク環境から新規または更新ドキュメントを収集し、データベース部１２に登録する。そして、その登録後に再び未分類のドキュメントを自動分類するために、分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５が起動される。ここで、ドキュメント収集部６６および分類情報抽出部６７の２回目以降の動作終了時に行われる自動分類処理では、分類特徴計算部１３は必ずしも起動する必要はなく、前回求められた各分類項目の特徴量を使って自動分類するようにしても良い。

【００７５】図１３は本発明の分散ドキュメント分類システムを実現するハードウェアの一例を示すブロック図であり、ＣＰＵ，メモリ，磁気ディスク，ディスプレイ装置，入力装置および通信装置等を含むデータ処理装置（コンピュータ）７１と、分散ドキュメント分類用プログラムを記録した記録媒体７２とから構成されている。記録媒体７２は、ＣＤＲＯＭ，半導体メモリ，磁気ディスクその他の記録媒体であって良い。分散ドキュメント分類用プログラムは記録媒体７２からデータ処理装置７１に読み込まれ、データ処理装置７１の動作を制御することにより、前述した第１の実施例にあっては、データ処理装置７１上に、ドキュメント手動登録部１１，データベース部１２，分類特徴計算部１３，ドキュメント特徴計算部１４および分類判定部１５を実現し、第２の実施例にあっては、ドキュメント手動登録部１１，データベース部１２，分類特徴計算部１３，ドキュメント特徴計算部１４，分類判定部１５，ドキュメント収集部６６および分類情報抽出部６７を実現し、第３の実施例にあっては、データベース部１２，分類特徴計算部１３，ドキュメント特徴計算部１４，分類判定部１５，ドキュメント収集部６６および分類情報抽出部６７を実現する。

【００７６】以上、本発明について幾つかの実施例を挙げて説明したが、本発明は以上の実施例にのみ限定されず、その他各種の付加変更が可能である。例えば、書誌項目として前述した具体例に示される以外のものを使用しても良く、タイトル等の付随情報を含ませても良い。また各分類項目の特徴量や各ドキュメントの特徴量の求め方も前述した具体例以外の方法を適用することが可能である。更に、分類項目はフラットな構造である必要はなく、階層構造を持った分類項目を使用することもできる。

【００７７】
【発明の効果】以上説明したように本発明によれば以下のような効果を得ることができる。
【００７８】ネットワーク環境に分散して存在するドキュメントの一部を手動で分類する作業をサービス提供者が行えば、その他のドキュメントは既に分類されているドキュメント群との類似度を計算して自動的に分類することができる。特に、手動登録においては、個々のドキュメントの書誌項目を調べてそれをドキュメント識別子と共にドキュメント情報格納部に登録する作業と並行して、そのように調査したドキュメントを実際に分類する作業が行える。このため、後の自動登録の際に使用する各分類項目ごとの特徴量の基礎となる書誌項目の設定を正確に行えるばかりか、一部のドキュメントの分類も同時に行えてしまうので無駄がない。
【００７９】収集されたドキュメント中に記述された分類項目指定を識別して分類する分類情報抽出部を備えることによって、ドキュメント作成者自身が分類を指定することができ、より正確な分類が可能となる。また、分類項目指定のなかったドキュメントについても自動分類することができる。
【図面の簡単な説明】
【図１】本発明の第１の実施例にかかる分散ドキュメント分類システムの構成を示すブロック図である。
【図２】分類情報格納部に格納される情報のデータ構造の一例を示すテーブルである。
【図３】ドキュメント情報格納部に格納される情報のデータ構造の一例を示すテーブルである。
【図４】分類特徴計算部の処理例を示すフローチャートである。
＊【図５】ドキュメント特徴計算部および分類判定部の処理例を示すフローチャートである。
【図６】ドキュメントの書誌項目および特徴量の具体例を示す図である。
【図７】分類項目の特徴量の具体例を示す図である。
【図８】未分類ドキュメントの書誌項目および特徴量の具体例を示す図である。
【図９】本発明の第２の実施例にかかる分散ドキュメント分類システムの構成を示すブロック図である。
【図１０】ドキュメント収集部および分類情報抽出部の処理例を示すフローチャートである。
【図１１】ドキュメント作成者がドキュメント中に記述する分類項目指定の例を示す図である。
【図１２】本発明の第３の実施例にかかる分散ドキュメント分類システムの構成を示すブロック図である。
【図１３】本発明の分散ドキュメント分類システムを実現するハードウェアの一例を示すブロック図である。
【図１４】従来の分散ドキュメント分類システムの構成を示すブロック図である。
【符号の説明】
１１…ドキュメント手動登録部
１２…データベース部
１２１…分類情報格納部
１２２…ドキュメント情報格納部
１３…分類特徴計算部
１４…ドキュメント特徴計算部
１５…分類判定部
６６…ドキュメント収集部
６７…分類情報抽出部

【図１】



【図２】

20 (121)

分類ＩＤ	分類	ドキュメント識別子リスト
class0	スポーツ	node001, node002
class1	芸能	node002, node004, node006
class2	社会	node011, node023, node027
class3	歴史	node022, node034, node036
class4	文学	node023, node025
・	・	・
・	・	・
・	・	・

【図３】

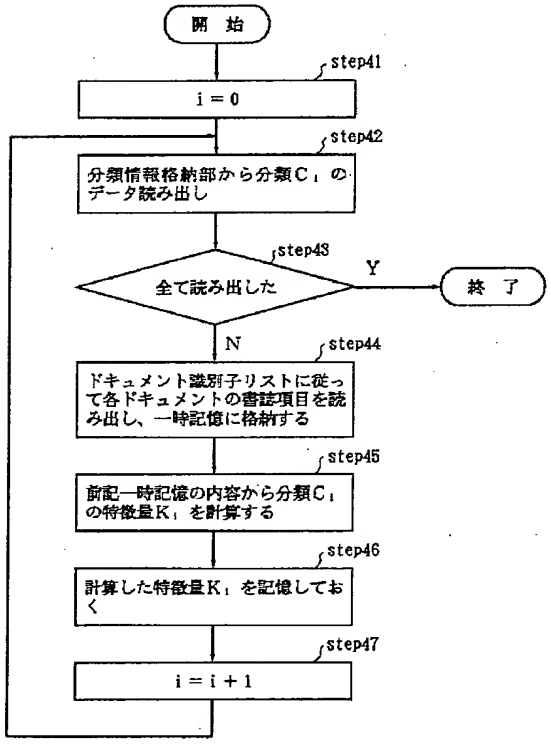
30 (122)

ドキュメント識別子	アドレス	書誌項目	分類方法
node001	host1.AAA.co.jp/DIR1/file1	titleA, key1, key3, key4	manual
node002	host2.AAA.co.jp/DIR1/file2	titleB, key2, key4	manual
node003	host1.BBB.co.jp/DIR2/file3	titleC, key4, key6, key7	auto
node004	host3.BBB.co.jp/DIR3/file4	titleD, key2, key5	auto
node005	host30.AAA.co.jp/DIR4/subdir1/file1	titleE, key5, key6, key7	auto
・	・	・	・
・	・	・	・
・	・	・	・

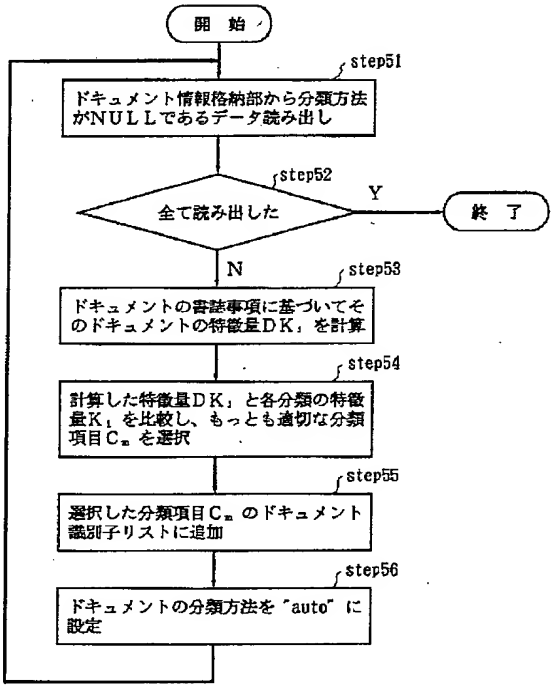
【図７】

単語	出現回数	出現頻度
テニス	8	8／23
スキー	10	10／23
サッカー	8	8／23
野球	2	2／23

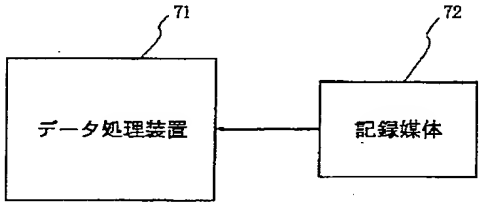
【図４】



【図５】



【図１３】



【図６】

(a)

単語	出現回数	出現頻度
テニス	5	5／12
スキー	4	4／12
サッカー	3	3／12

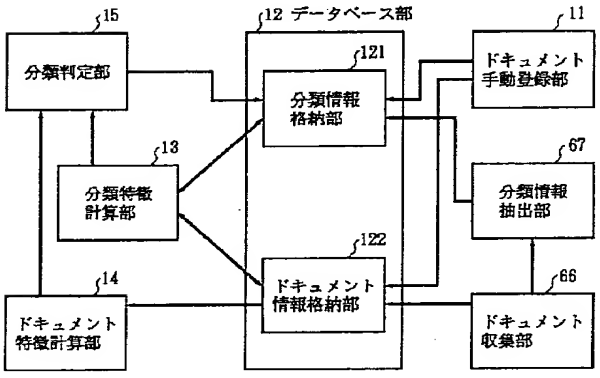
(b)

単語	出現回数	出現頻度
テニス	3	3／11
野球	2	2／11
スキー	6	6／11

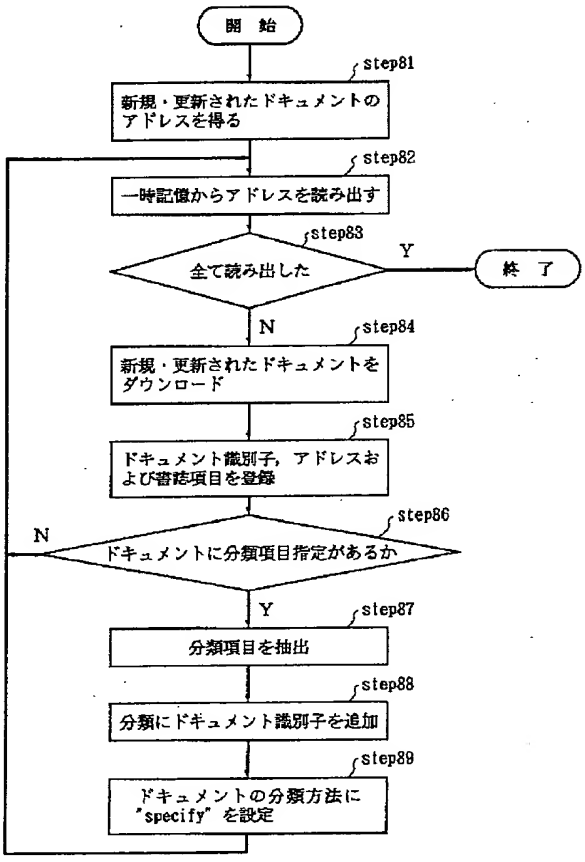
【図８】

単語	出現回数	出現頻度
テニス	4	4／9
スキー	2	2／9
ゴルフ	3	3／9

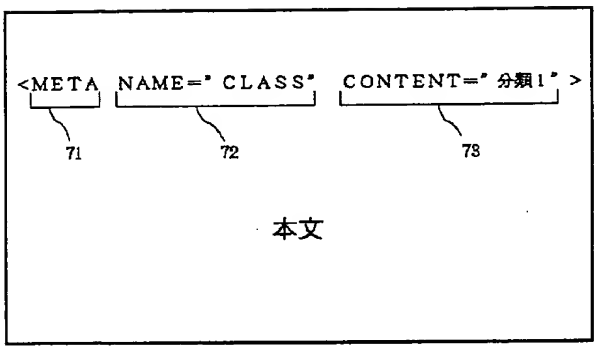
【図９】



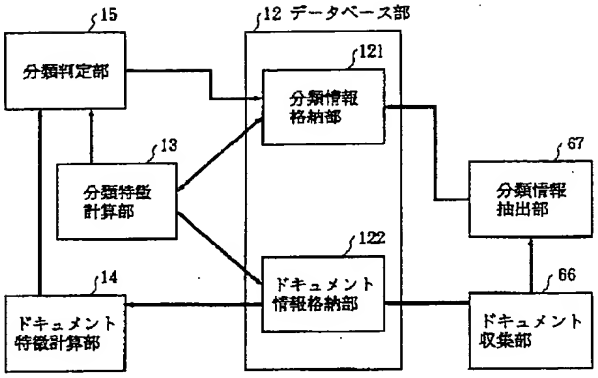
【図１０】



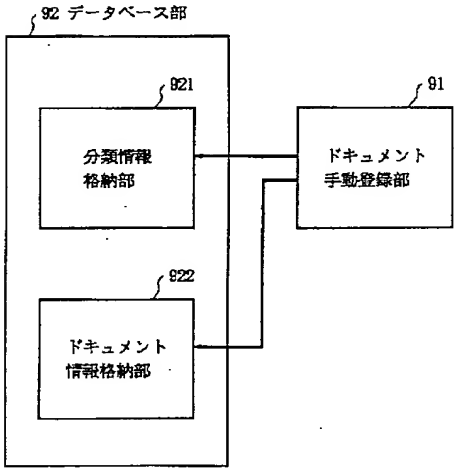
【図１１】


<META NAME=" CLASS" CONTENT=" 分類1" >
71
72
73
本文

【図１２】


15
分類判定部
12 データベース部
121
分類情報格納部
13
分類特徴計算部
67
分類情報抽出部
122
ドキュメント情報格納部
14
ドキュメント特徴計算部
66
ドキュメント収集部

【図１４】


92 データベース部
921
分類情報格納部
91
ドキュメント手動登録部
922
ドキュメント情報格納部